DEIM Forum 2014 F9-5

画像の特徴抽出手法を利用した楽曲のセグメンテーション

三浦　亮†　喜田　拓也†

† 北海道大学大学院情報科学研究科　〒 060–0814 北海道札幌市北区北 14 条西 9 丁目
E-mail: †{ryo_miura,kida}@ist.hokudai.ac.jp

あらまし　ポピュラー音楽と呼ばれるジャンルの音楽には，イントロやサビといった音楽的に意味を持つ構造が存在する．現在も増加し続ける音楽データを効率よく管理するためには，こうした楽曲構造を理解することが重要である．また楽曲構造を解析し抽出することによって，より高度な楽曲鑑賞を行うことができる．本稿では，楽曲信号から得られる自己相似性行列に潜在するブロック構造を抽出することで，楽曲の構造の切れ目を発見する手法について議論する．その際，入力された楽曲の音高と音色それぞれを特徴量とした二種類の自己相似性行列を用いる．それらを二次元画像とみなし，数種類の画像フィルタリングと特徴構造を抽出するためのハフ変換を利用して楽曲の構造を割り出す手法を提案する．

キーワード　情報処理，楽曲構造解析，画像フィルタリング，特徴抽出，サビ抽出，楽曲理解，ハフ変換

1. はじめに

昨今、インターネット上で利用可能なデジタル音楽データの数が増加し続けている。たとえば，iTunes Store では 2600 万を超える数の楽曲が配信されており(注1) 利用者はいつでも自分の好きな楽曲購入し楽しむことができる．

ユーザ生成コンテンツ (UGC) の成長も著しい．配信されている楽曲を購入するのみならず，自らが制作した楽曲を YouTube などのコンテンツ共有サービスへ投稿し，全世界へ公開する人が増えた(注2)．音楽制作ソフトウェアの性能向上と普及により，オリジナルの楽曲を制作するアマチュアが増えている．ヤマハ株式会社によって開発された VOCALOID(注3) の誕生は UGC 発展の要因のひとつである．

そのような潮流の中，楽曲が持つ構造はこのようなオンライン上に存在するデータを分類し整理する上で重要な役割を果たしている．たとえば，イントロやサビと呼ばれるものが音楽構造に該当する [9]．楽曲試聴を行う上で，曲の開始時点のみからの再生以外にも，サビからの再生などができると，試聴におけるバリエーションが増え便利である．実際，楽曲試聴を行う際に，サビからの再生ができる再生機，試聴機も存在する．これは各楽曲にセグメンテーション・メタデータ（セグメントタグ）を添加することによって行うことができる．また，このようなメタデータは，ユーザーに適切な曲を推薦するための検索時に重要な手がかりを提供する．クラスタリングを行うよりも効果的に楽曲検索を可能にする [5] など，大規模な音楽データを管理するシステムのために有用である．

構造の解析に使われるデータ構造として存在する手法として，自己相似性行列 (SSM) [6] という構造を用いるものがある．楽曲を入力データとした際に，SSM は楽曲構造を表し，これを可視化できるデータ構造である (図 1)．行列の各要素は，音楽の一部分に対応する部分シーケンス間の類似度を表す．この SSM 上に存在する同じような構造をとる領域を見つけることで，楽曲の構造を決定することができる．

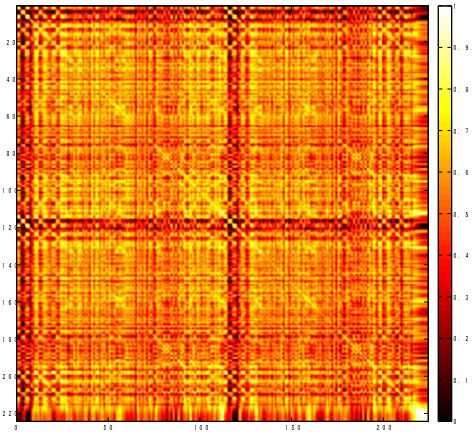

図 1　自己相似性行列 (Self-Similarity Matrix) の例 : 入力データには RWC 研究用音楽データベースのクラシック音楽データベース RWC-MDB-C-2001-C04，楽曲番号 03 :「シューベルト - ピアノ五重奏曲イ長調 op.32 D.667 『ます』 第 4 楽章」を用いた．

本稿では，SSM に対して輪郭抽出画像フィルタリングを行い，さらにハフ変換による画像内特徴構造抽出をすることで楽曲のセグメンテーションを行い，楽曲構造を解析する手法を提案する．楽曲のセグメンテーションとは，ある楽曲について，サビ，イントロ，A メロなどというような音楽的な意味を持つ部分に分割することを意味する．楽曲構造を推定するための SSM を構築するにあたり，SSM を計算するための特徴量を決める必要がある．楽曲の解析では，音の高低を表す特徴量であるクロマベクトル，また本稿の実験では音色に関する特徴量であるメル周波数ケプストラム (MFCC) についての類似度の計

(注1) : http://www.apple.com/jp/itunes/
(注2) : http://www.youtube.com/
(注3) : http://www.vocaloid.com/en/

算が用いられる [4]．本稿の手法は，推定される楽曲構造の区切れ目を的確に発見し，[18] で採られている手法よりも高い精度での解析を実現した．

ハフ変換を用いた楽曲構造解析手法には [1] のようなものがある．この手法は SSM に存在する画像に対して斜めに走る直線構造を抽出し，解析に利用するものである．この手法とは異なり，本稿で議論する手法は SSM に対して縦方向に存在する構造を抽出するものである．

2. 音 声 処 理

2.1 楽曲が持つ構造

一般的に，それぞれの楽曲は独自かつ独特の構造を有している．楽曲には区間というものがあり，区間はある程度の長さを持つ構造である．これらの区間から楽曲が成り立っていると考えることもできる．一方で，楽曲の譜面には小節という区切り方がある．小節は特定の拍節数で定義された構造の区切り方であり，特定の音符の値が与えられる．区間は複数の小節からなることが多い．

ボーカルの有無を問わず，イントロ，ブリッジ，コーラスなどのような特定の音調や音楽的な意味を持っている部分は区間となる．それぞれの楽曲でセクション長が異なる場合でも，すべての楽曲は分割が可能である．楽曲構造解析は，各区間が繰りかえされる構造を発見することで行われる [14]．

後藤 [9] は，音楽音声信号を入力とした楽曲構造解析の手法を提案し，サビ区間自動抽出を行う音楽試聴インタフェースである “SmartMusicKIOSK” を示した．この手法では楽曲構造解析をクロマベクトル間の類似度の計算を通じて行う [4]．この研究は “Songle” [11] として Web サービスとして結実し，現在 Web 上でこのインタフェースを使用することができる．Paulus と Klapuri [18] は，確率的適合指標を利用する別の楽曲の区間分け手法を提案した．彼らの手法では，音の高低を示す特徴量であるクロマベクトル，音色の異なりを表す特徴量であるメル周波数ケプストラム係数 (MFCC)，および楽曲のリズム・拍節の音を表すリズモグラムを利用して楽曲構造解析を行う．楽曲の区間分けを行う際，手法 [9] と [18] 双方は，自己相似性行列を計算する．

2.2 クロマベクトル

クロマベクトル [7] は 12 次元の音声信号特徴量であり，極めて短い断片においてそれぞれの音階を周波数軸としたパワー分布を表す特徴量である．音楽信号を非常に短く区切り，区切られたそれぞれの断片に対して周波数スペクトルについてフーリエ変換を行う．こうして計算されたスペクトルパワーを 12 半音階ごとに加算し蓄積する．

ある節回しが楽曲中で繰り返すされる際，その楽曲が持つ旋律や伴奏が少なからず変化をしながら繰り返されることがある．しかし，全体の響き (同時になっているすべての音の構成) が類似しているならば，繰り返しの区間として検出することが可能になる．たとえば，サビ部分の歌詞が 1 番と 2 番で異なっていてもメロディーが同様ならば同じものとして扱うようなものに近い．そのような構造を高い精度で検出することが期待できる構造である．

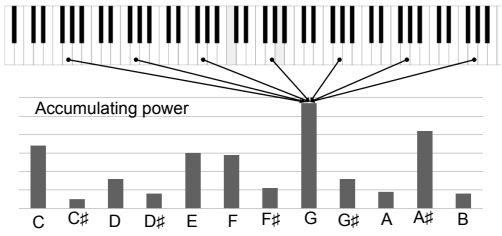

図 2　クロマベクトル算出の仕組み．

2.3 メル周波数ケプストラム (MFCC)

ケプストラムとは，音のスペクトルを信号と見なし，音の振幅スペクトルに対数をかけ，さらにフーリエ変換を行ったもの，つまりスペクトルのスペクトルのことををケプストラムという．この処理を行うことで，音源の持つ周波数と声道の持つ周波数を切り分けることができる．音源の持つ周波数とは基本周波数のことであり，この対数スペクトルの高次成分にあたる．また声道の持つ周波数とは共鳴音の周波数を指し，こちらは低次成分にあたる．この対数ケプストラムの低次成分である声道の持つ周波数特性に対して，人間が持っている周波数を知覚する特性を加味した重み付けを行った特徴量を MFCC [13] という．

音の高さの知覚特性から得られた尺度のことをメル尺度という．MFCC は，このメル尺度を用いて声道の周波数特性を変換したものである．メル尺度とは，1000Hz の音を 1000 メルの音高とし定義し，この基準音から 2 倍の高さの音を 2000 メル，半分の高さの音を 500 メルとするものである．

2.4 自己相似性行列 (Self-similarity Matrix)

入力する楽曲の音声データを m 個の断片に分割する．以降では，それぞれの断片をフレームと呼ぶ．i 番目のフレームのクロマベクトルをパワースペクトルについてノーマライズし，それを $\mathcal{X}_i = (\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{12}^{(i)})$ とすると，i 番目のフレームと j 番目のフレームの区間類似度は，

$$s_{i,j} = \frac{\langle \mathcal{X}_i, \mathcal{X}_j \rangle}{|\mathcal{X}_i| \cdot |\mathcal{X}_j|}.$$

で求められる．

区間類似度は，何らかのシーケンスを短く区切った断片が他の断片と比較してどれだけ類似しているかを表す尺度である．楽曲構造解析で言えば，シーケンスは解析を行う対象となる楽曲であり，比較を行う対象はその断片でどのような音が鳴っているかである．自己相似性行列 (Self-similarity Matrix, SSM) [6], [15], [17], [19] は，すべての断片の組み合わせの区間類似度を可視化することができる．

区間類似度を行列状に配置したものが SSM となる．縦方向，横方向，ともに短い時間の区切りを単位とする．楽曲が m 個のフレームに分けられているとき，SSM は m 次正方行列状となり，SSM の i 行 j 列目の要素は i フレーム目と j フレーム目の類似度が挿入される．このとき，数値の高低を画像の明暗で塗り分けることにより，その楽曲中で類似した部分が行列構造に帯状となって現れる [15], [16]．したがって，楽曲の中で繰り返

して使われているフレーズがどのあたりに存在し，そのフレーズがどれくらいの時間演奏されているか，ということが視覚的にも理解できるようになる．

類似性は，空間的な距離や，相関性などの比較によって説明することができる．SSM を用いる例としては，遺伝子配列のマッチングのようにデータ系列によって指定されたパターンの検索 [21] や，人間の動作の類似性の計測 [12]，またオーディオ信号中における似た音の探索 [8] などがある．

3. 画 像 処 理

3.1 輪郭抽出フィルタ

デジタル画像処理におけるフィルタリング技術のひとつに，領域境界抽出と呼ばれるものがある [20]．領域境界抽出とは，対象となる画像について，画素の明度が急激に変化する点を抽出するものである．画像において明度が急激に変化する箇所とは画像に存在する構造の輪郭を表すものであるため，領域境界抽出は言い換えると画像に含まれている構造の輪郭を抽出するものであるとも言える．明度の変化は微分計算を利用することで算出する．

輪郭抽出フィルタの例として，Scharr フィルタを上げる．Scharr フィルタは領域境界抽出に使用される一次微分で勾配を計算するフィルタである．画像の縦方向に潜在する輪郭を抽出するためのカーネルは以下のようになる．

$$\begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}.$$

入力画像のすべての画素について，このカーネルを使用して近傍領域との差分を計算すると領域境界を抽出できる．フィルタリング行列を上式で表される行列とし，さらに入力として与えるデジタル画像を I，出力されるデジタル画像を R とする．このとき，出力画像のある画素 $R(x,y)$ は入力画像の画素 $I(x,y)$ から $R(x,y) = -3(I(x-1,y-1)+I(x-1,y+1))-10(I(x-1,y))+10(I(x+1,y))+3(I(x+1,y-1)+I(x+1,y+1))$ で求められる．このカーネルを転置して計算を行うと，画像に対して横方向の領域境界を抽出するようになる．

このような輪郭抽出を行う画像フィルタリングを SSM に行うことで，続いて説明するハフ変換を用いた特徴構造抽出・解析の補助を行うことができる．

また，本稿における実験では，Scharr フィルタと同様に画像の輪郭抽出を行うための画像フィルタリング技術として，Sobel フィルタと Canny フィルタを比較に用いている．Sobel フィルタは，Scharr フィルタと同様に空間 1 次微分を計算し輪郭抽出を行うフィルタであり，以下のように表される．

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}.$$

Canny フィルタは，輪郭抽出の前に画像全体にガウスフィルタをかけ画像全体を平滑化し，その後輪郭抽出を行う 1 次微分フィルタをかけるものである．

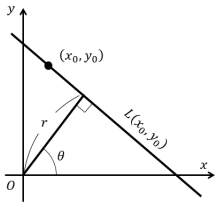

図 3　平面に存在する直線をハフ平面で扱うための準備．(x,y) による表現を (r,θ) による表現へ変換する．

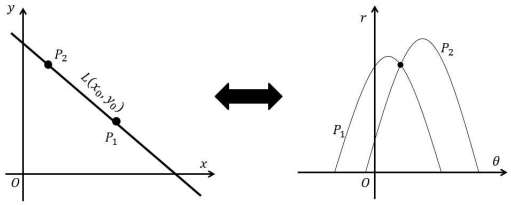

図 4　$x-y$ 平面の直線をハフ平面に再配置する．

3.2 ハ フ 変 換

ハフ変換 [2] は，デジタル画像処理における特徴抽出の手法である．とくに画像内に潜在する直線構造を検出し抽出するものである．パラメータ空間への投票に基づく特徴抽出手法である．

平面上に存在するある 1 点 P を考える．点 P を通る平面状の直線は無数に存在する．この直線の集合の中で，平面，あるいは画像における特徴点，つまり画像に含まれている構造の輪郭に当たる部分を指す点を通過するような直線を探索する．結果として抽出されたものが，画像における特徴的な直線となる．

ハフ変換では，1 つの直線を 2 つのパラメータ r,θ で表す．この 2 つのパラメータで表現されるようなパラメータ空間において特徴抽出を行うものである．ハフ変換では，原点 O から対象となるある 1 本の直線に下ろした法線の長さ r と，その角度 θ を用いて，その直線が通っている点 $\mathrm{P}(x,y)$ の座標を用いて，$r = x\cos\theta + y\sin\theta$ と表すことができる．よって，任意の直線に対する (r,θ) の組み合わせは一意に定めることができる．このとき r と θ からなる平面はハフ平面 (図 4) と呼ばれる．

解析を行う画像について，その画像の特徴点と思われる点を通過しているすべての直線について，(r,θ) のパラメータ表記化する．パラメータ表記化された直線をハフ平面にプロットすると，ハフ平面は特徴点を通過する直線を表す正弦曲線で満たされる．ハフ平面上の正弦曲線が重なっている点を検査していき，重なっている箇所に「投票」を行う．ハフ平面上で正弦曲線が交差している点は，元の画像において特徴点とされる点を多数通過している直線構造となり，これが元の画像の特徴構造となる．すべての投票を終えた結果，数多くの交差が確認された点を上位から並べて，これを元画像の特徴構造として出力する．したがって，この「投票」の結果から元の画像に潜在する特徴的な直線構造を検出することができる．

4. 提 案 手 法

この節では，本稿の楽曲構造解析手法について説明する．

本稿の処理の流れは以下のとおりである．本稿において，SSM 画像における特徴構造の抽出は，入力された楽曲データのセグメンテーションを行うことと同義である．最初に，構造解析を行う楽曲の音声データを入力し，楽曲の音高遷移を示すクロマベクトルと楽曲の音色遷移を示す MFCC を計算する．次いで，計算されたクロマベクトル列と MFCC 列からそれぞれ SSM を計算する．また，音高遷移と音色遷移の双方を計測したものと同義となるように，それぞれの特徴量からなる 2 つの SSM を組み合わせた SSM も準備する．本稿ではクロマベクトルからなる SSM と MFCC からなる SSM の和と積を用意した．計算された SSM に対して，画像の輪郭抽出フィルタをかけ SSM に潜在する特徴構造の顕在化を行う．以上のように生成した SSM，ならびに輪郭抽出のを行った SSM に対してハフ変換を行うことで，画像の特徴構造の抽出を行うというものである．

入力された音声データを解析しクロマベクトルへ変換する．すなわち，楽曲データを 12 次元 (ドから半音階刻みでシまで) ベクトルの列からなるデータに変換する．さらにこのデータについて正規化を行ってノイズとなる成分を除去する．また音色についても同様に解析する．楽曲データを MFCC へ変換し，12 次元のデータ列に変換する．

図 5 では，クロマベクトルによる SSM の例を示している．例には RWC-MDB-P-2001 No.7 『PROLOGUE』(注4) を使用した．図 6 では MFCC を用いた SSM の例を示している．

図の左上が楽曲データの開始地点，右下が終了地点を表し，楽曲は右下へ向かって進行する．SSM を構成する各要素はフレーム間の音声の類似度を表し，類似度が高い箇所ほど明るい色で表示されている．類似度が高い箇所がある程度の長さを持って現れるような楽曲では，SSM を計算した際に明度の高いブロックのような正方形構造が現れるこれが楽曲構造である．楽曲におけるサビなどの区間において，その区間内で鳴っている音は似ていることが多いため，SSM で明度の高いブロック構造が現れてきた箇所を楽曲の区間として考えることができる．

図 5 の SSM に対して Scharr フィルタによるフィルタリングを施したものが図 7 である．縦の輪郭を抽出するような Scharr フィルタで処理を行うことにより，SSM におけるブロック構造の始点部分の輪郭を抽出することができる．この処理は，次の処理であるハフ変換による特徴的直線構造を判別に有用である．

また，クロマベクトルからなる SSM と MFCC からなる SSM を統合したものも用いる．例として，2 つの SSM の和 (図 8) と積 (図 9)

これら数パターンの SSM に対してハフ変換を行う．図 10 は，図 5 に対して，図 11 は，図 7 に対して，それぞれハフ変換を行ったものである．ハフ変換を実行した結果検出された直線構造の中から，画像に対して正しく垂直になっている直線構造のみを選び出し，その座標を記録していく．SSM から検出され抽出された座標の値は，入力された楽曲の再生時刻を指す．これらの座標の値をまとめた結果が，本稿の解析手法によって楽曲構造の区切れ目と扱われた部分となる．

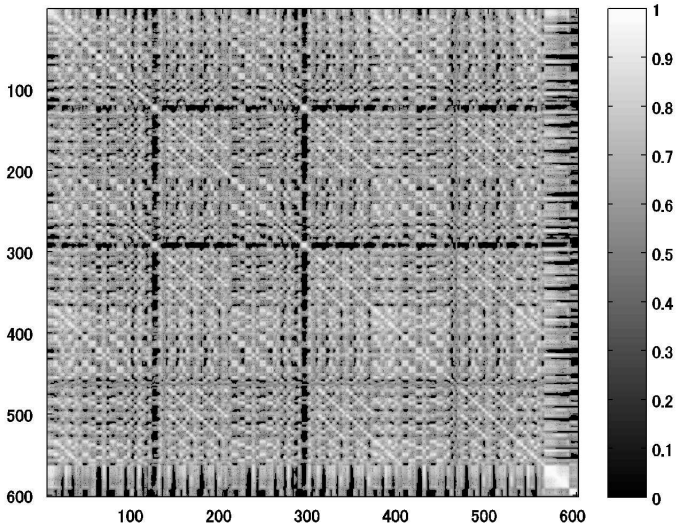

図 5　クロマを用いた自己相似性行列 (SSM) の例.

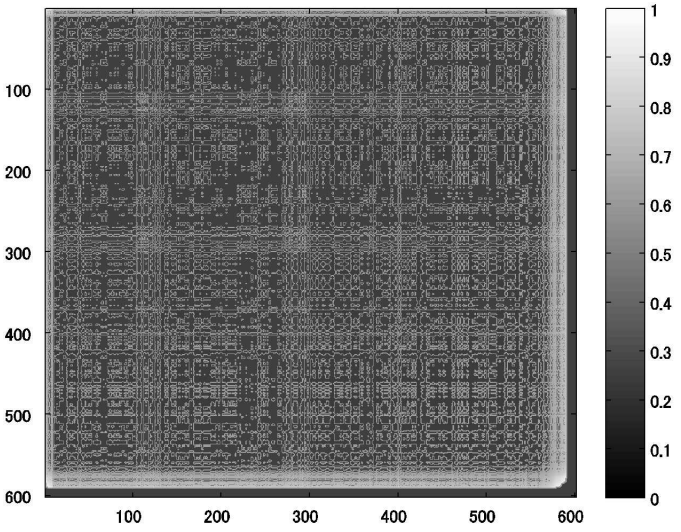

図 6　MFCC を用いた自己相似性行列 (SSM) の例.

また本稿の実験では，転調を考慮した SSM と考慮しない SSM の双方を用いる．転調を考慮した SSM は，ベクトルの回転演算を次元の数だけ行い，回転させたベクトルと元のベクトルとの類似度を計算していくことで計算する．

5. 実　　験

楽曲構造の境界判定能力の比較実験を行った．実験には，RWC 研究用音楽データベースのポピュラー音楽全 100 曲(注5) を使用した．また楽曲構造の Ground Truth となるデータには，RWC 音楽データベースのための AIST Annotation データから，ポピュラー音楽データベースの Chorus Section アノテーションデータを使用した [10]．このデータは，RWC 研究用音楽データベースに収録されている楽曲について，楽曲のサビ区間や A メロ，あるいは前奏，間奏などの区間が，それぞれどこからどこまでであるのか，ということを記録した楽曲構造解析

(注4)：演奏時間 5 分 00 秒

(注5)：RWC-MDB-P-2001 No.01-100

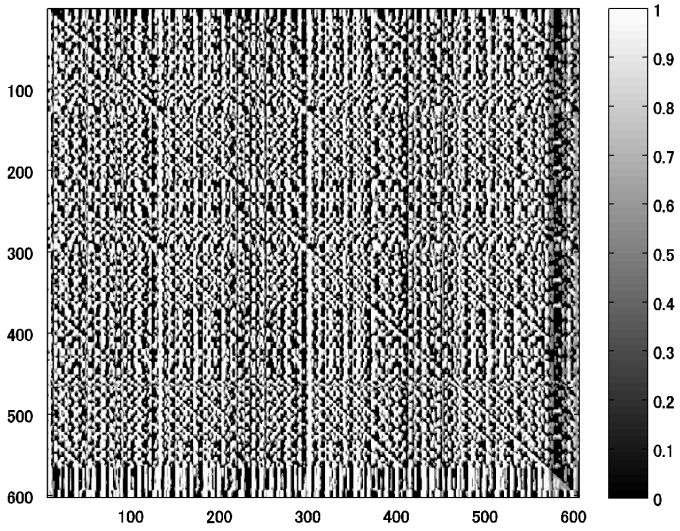

図 7　図 5 に対して Scharr フィルタによる処理を行ったもの.

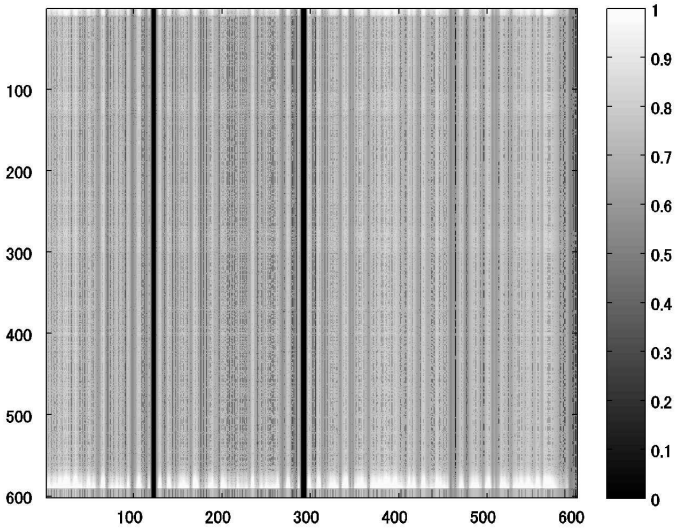

図 9　2 つの SSM を乗算したもの.

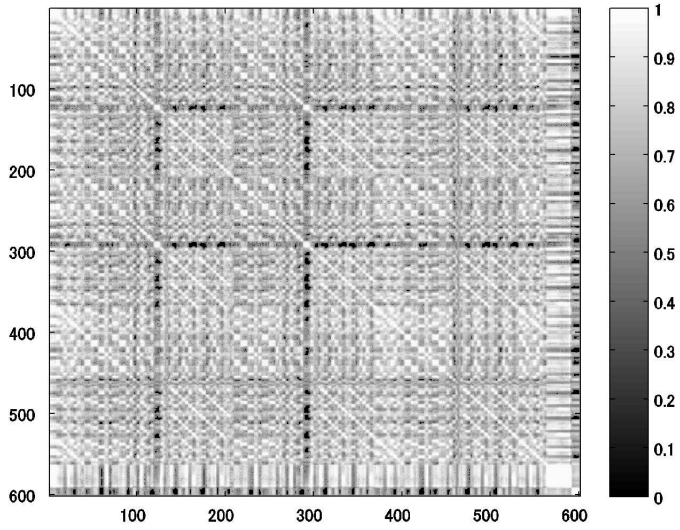

図 8　2 つの SSM を和算したもの.

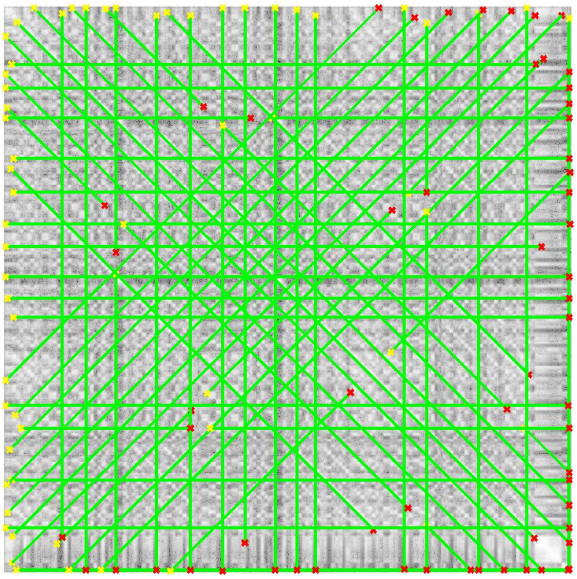

図 10　図 5 に対してハフ変換を行ったもの.

のための補足資料である.

比較実験の内容は，楽曲構造解析の結果得られた構造の区切れ目と思われる楽曲再生経過秒数のデータと，AIST Annotation データで記されている正解とする区切れ目の時刻データの一致度を調べるものである.

最初に，区間の区切れ目に対して，楽曲全体についてクロマベクトルの列と MFCC の列をそれぞれ前から順にクラスタ分割を行う．クラスタへの分割の例をあげる．楽曲の特徴量ベクトルが 8 本あり，これを解析した結果として 3 番目と 5 番目に区切れ目がある場合を考える．このとき，各特徴量ベクトルに対して，$\begin{pmatrix}1 & 1 & 2 & 2 & 3 & 3 & 3 & 3\end{pmatrix}$ とラベリングして 3 つのクラスタを生成するものとする．ラベル付けによって，境界で区切られた区間が特徴量ベクトル列上ではどの範囲であるかを示すことができるようになる.

精度の評価指標としては，調整ランド指数 (adjusted rand index) [3], [22] を用いる．調整ランド指数は，2 つのクラスタリングの類似度を見るために用いられる指標である．指数の値は-1 から 1 の間をとり，クラスタのラベリングが完全に一致していれば 1 となる．n 個の要素からなる集合 S と，2 つの比較したいクラスタリングを与える．この 2 つの配列をそれぞれ $X = X_1, X_2, \cdots, X_r$, $Y = Y_1, Y_2, ..., Y_s$ とする．さらに，2 つの配列の要素 X_i, X_j について，$n_{ij} = |X_i \cap Y_j|$ をすべての組み合わせについて計算し，この値について行，列ごとに総和を取る．行ごとの総和を $a_1, a_2, ..., a_r$, 列ごとの総和を $b_1, b_2, ..., b_s$ とする.

以上の値を用いて，調整ランド指数 ARI は，

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

と定義される．配列の片方に Ground Truth データを代入することで，クラスタリングの正確さを評価できる.

上記の方法で，RWC 研究用音楽データベースのポップス楽曲全 100 曲を用いて実験を行った．本稿では，楽曲データをクロマベクトルと MFCC に変換する際に，0.5 秒毎に 1 サンプル

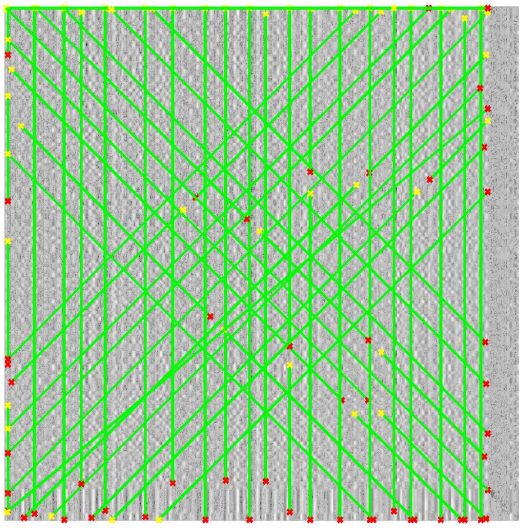

図 11　図 7 に対してハフ変換を行ったもの.

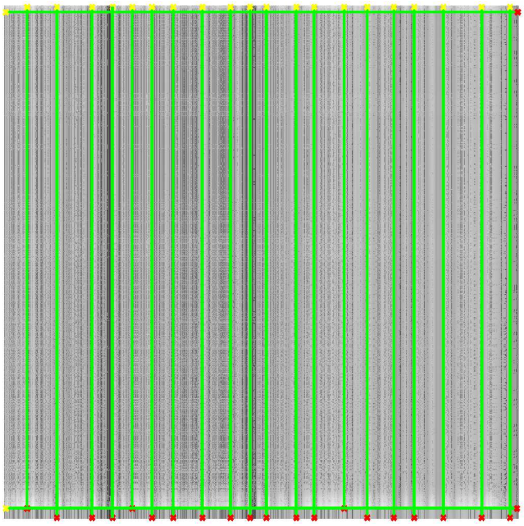

図 12　図 9 に対してハフ変換を行ったもの.

を取得して算出した低解像度の SSM と，0.2 秒毎に 1 サンプルを取得して算出した高解像度の SSM を用意した．高解像度の SSM に対しては，画像全体に先んじてガウスフィルタをかけて平滑化を行っている．また，それぞれ以下の 71 通りの処理の組み合わせに対して行った上で調整ランド指数を計算した．

（1） SSM に Hough 変換を施して ARI 算出 (Normal)

（2） 転調を考慮した SSM に Hough 変換を施して ARI 算出 (Normal-Mod)

（3） SSM に Scharr フィルタをかけ Hough 変換を施して ARI 算出 (Scharr)

（4） SSM に Sobel フィルタをかけ Hough 変換を施して ARI 算出 (Sobel)

（5） 転調を考慮した SSM に Sobel フィルタをかけ Hough 変換を施して ARI 算出 (Sobel-Mod)

（6） 転調を考慮した SSM に Scharr フィルタをかけ Hough 変換を施して ARI 算出 (Scharr-Mod)

（7） SSM に Canny フィルタをかけ Hough 変換を施して ARI 算出 (Canny)

表 1　低解像度 SSM を用いた解析の結果.

ARI 値	Average	Variance
(1)Normal	0.552343	0.003778
(2)Normal-Mod	0.548928	0.006084
(3)Scharr	0.562042	0.002493
(4)Scharr-Mod	0.577466	0.003778
(5)Sobel	0.563904	0.002971
(6)Sobel-Mod	0.575256	0.004088
(7)Canny	0.475407	0.019036
(8)Canny-Mod	0.467464	0.012060
(9)MFCC	0.498812	0.014477
(10)Mix	0.561228	0.002999
(11)Mix-Mod	**0.581846**	0.003539
(12)Add(-Mod)	0.540614	0.005963
(13)MFCC+Scharr	0.566774	0.003570
(14)Mix+Scharr	0.546201	0.002361
(15)Mix+Scharr-Mod	0.528123	0.002389
(16)Add+Scharr(-Mod)	0.556774	0.002444
(17)Gaussian	0.478079	0.009000

（8） 転調を考慮した SSM に Canny フィルタをかけ Hough 変換を施して ARI 算出 (Canny-Mod)

（9） MFCC で算出した SSM に Hough 変換を施して ARI 算出 (MFCC)

（10） 2 つの特徴量でそれぞれ生成した SSM を乗算し，Hough 変換を施して ARI 算出 (Mix)

（11） 転調を考慮した上で 2 つの特徴量でそれぞれ生成した SSM を乗算し，Hough 変換を施して ARI 算出 (Mix-Mod)

（12） 2 つの特徴量でそれぞれ生成した SSM を和算し，Hough 変換を施して ARI 算出 (Add(-Mod))

（13） MFCC で算出した SSM に Scharr フィルタをかけ Hough 変換を施して ARI 算出 (MFCC+Scharr)

（14） 2 つの特徴量でそれぞれ生成した SSM を乗算し，それに Scharr フィルタをかけ Hough 変換を施して ARI 算出 ((Mix+Scharr)

（15） 転調を考慮した上で 2 つの特徴量でそれぞれ生成した SSM を乗算し，それに Scharr フィルタをかけ Hough 変換を施して ARI 算出 (Mix+Scharr-Mod)

（16） 2 つの特徴量でそれぞれ生成した SSM を和算し，それに Scharr フィルタをかけ Hough 変換を施して ARI 算出 (Add+Scharr(-Mod))

（17） [18] で用いられているガウシアンフィルタを応用したフィルタリング手法の一部 (既存手法) から ARI 算出 (Gaussian)

計算結果を表 1 ならびに表 2 に示す．比較の結果，ハフ変換を用いた解析手法は概ね Gaussian と比較して良好な値を示していることがわかる．これは比較手法と比べ，本稿の手法での解析は正解データに近い区切れ目数に基づいたセグメンテーションを行えている場合が多いためと考えられる．とくに，クロマベクトルの SSM と MFCC の SSM を乗算した SSM を用いた際に，SSM の解像度に依らず優れたセグメンテーションが行えていることがわかる．また転調を考慮した SSM を用い

表 2　高解像度 SSM を用いた解析の結果.

ARI 値	Average	Variance
(1)Normal	0.458862	0.016489
(2)Normal-Mod	0.345133	0.028130
(3)Scharr	0.526009	0.006350
(4)Scharr-Mod	0.462222	0.016903
(5)Sobel	0.464984	0.019494
(6)Sobel-Mod	0.455728	0.014472
(7)Canny	0.521692	0.007766
(8)Canny-Mod	0.467008	0.018997
(9)MFCC	0.491412	0.016021
(10)Mix	0.562405	0.003282
(11)Mix-Mod	**0.575249**	0.002825
(12)Add(-Mod)	0.399869	0.021194
(13)MFCC+Scharr	0.561439	0.004002
(14)Mix+Scharr	0.554166	0.003181
(15)Mix+Scharr-Mod	0.539761	0.002844
(16)Add+Scharr(-Mod)	0.456021	0.020913
(17)Gaussian	0.473289	0.008679

た方が転調を考慮しない SSM を用いた時より良い解析結果が得られ，この場合についても良いセグメンテーションが行えていることがわかる．また本稿の手法においては，必ずしも高解像度の SSM を用いたほうが高い精度での解析を行えるとは言えず，場合によっては解像度が低い SSM を用いたほうが精度の高い解析を行えていることもあった．低解像度 SSM の算出は，高解像度 SSM の算出よりも高速に実行できるため，本稿の手法ではマシンパワーが低いコンピュータでも楽曲の構造解析が行いやすいと言える．

6. まとめと今後の課題

本稿では楽曲構造の解析を行うために，クロマと MFCC の類似度から計算された自己相似性行列と，デジタル画像処理，とくに画像の輪郭抽出の手法とを組み合わせた手法について議論した．調整ランド指数というクラスタリング結果比較を行う指標を用いて解析結果を比較すると既存手法よりも良好な数値が得られた．また他の画像フィルタリング手法を輪郭抽出に用いた場合の実験も行った．

本稿における実験では，クロマから生成された SSM と MFCC から生成された SSM を掛けあわせて算出された SSM に対してハフ変換を行った際に最も良好なセグメンテーションが得られるという結果になった．この算出方法は，入力された楽曲の音の高さ，つまりメロディーの構造と，入力された楽曲の音色の双方を考慮していると言え，より楽曲の内容に迫った解析ができている．また解像度の異なる SSM に対してそれぞれの手法を適用した結果，必ずしも解像度を上げれば解析精度も高まるとは言えないことがわかった．

今後もより正確なクラスタリングを行うための機構について調査していきたい．

文　献

[1] Aucouturier, J.-J. and Sandler, M.: Finding Repeating Patterns in Acoustic Musical Signals: Applications for Audio Thumbnailing, *Proceedings of the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES22)*, pp. 412–421 (2002).

[2] Ballard, D. H.: Generalizing the Hough transform to detect arbitrary shapes, *Pattern Recognition*, Vol. 13, No. 2, pp. 111 – 122 (1981).

[3] Barrington, L., Chan, A. B. and Lanckriet, G.: Modeling Music as a Dynamic Texture, *IEEE Transactions on Audio. Speech, and Language Processing*, Vol. 18, No. 3, pp. 602–612 (2010).

[4] Bartsch, M. and Wakefield, G.: To catch a chorus: using chroma-based representations for audio thumbnailing, *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, pp. 15–18 (2001).

[5] Bartsch, M. A. and Wakefield, G. H.: Audio thumbnailing of popular music using chroma-based representations, *Multimedia, IEEE Transactions on*, Vol. 7, No. 1, pp. 96–104 (2005).

[6] Cooper, M. and Foote, J.: Summarizing popular music via structural similarity analysis, *Proceedings 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 127–130 (2003).

[7] Ellis, D. P. W.: Classifying music audio with timbral and chroma features, *ISMIR 2007: Proceedings of the 8th International Conference on Music Information Retrieval*, Austrian Computer Society, pp. 339–340 (2007).

[8] Foote, J.: Visualizing music and audio using self-similarity, *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, ACM, pp. 77–80 (1999).

[9] Goto, M.: A Chorus-Section Detecting Method for Musical Audio Signals, *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. V–437–440 (2003).

[10] Goto, M.: AIST Annotation for the RWC Music Database, *Proceedings of the 7th International Conference on Music Information Retrieval*, pp. 359–360 (2006).

[11] Goto, M., Ogata, J., Yoshii, K., Fujihara, H., Mauch, M. and Nakano, T.: PodCastle and Songle: Crowdsourcing-Based Web Services for Retrieval and Browsing of Speech and Music Content, pp. 36–41 (2012).

[12] Junejo, I. N., Dexter, E., Laptev, I. and Pérez, P.: Cross-View Action Recognition from Temporal Self-similarities, *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, Berlin, Heidelberg, Springer-Verlag, pp. 293–306 (2008).

[13] Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling, *International Symposium on Music Information Retrieval* (2000). Electronic Edition.

[14] Lu, L., Wang, M. and Zhang, H.-J.: Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data, *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 275–282 (2004).

[15] Müller, M. and Clausen, M.: Transposition-Invariant Self-Similarity Matrices, *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, pp. 47–50 (2007).

[16] Müller, M., Grosche, P. and Jiang, N.: A Segment-Based Fitness Measure for Capturing Repetitive Structures of Music Recordings, *Proceedings of 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, USA, ISMIR, pp. 615–620 (2011).

[17] Müller, M. and Kurth, F.: Towards structural analysis of audio recordings in the presence of musical variations, *EURASIP J. Appl. Signal Process.*, Vol. 2007, No. 1, pp. 163–163 (2007).

[18] Paulus, J. and Klapuri, A.: Music structure analysis using

a probabilistic fitness measure and a greedy search algorithm, *IEEE Transactions on Audio, Speech, and Laguage Processing*, Vol. 17, No. 6, pp. 1159–1170 (2009).

[19] Paulus, J., Müller, M. and Klapuri, A.: State of the Art Report: Audio-Based Music Structure Analysis, *Proceedings of the International Conference on Music Information Retrieval 2010* (2010).

[20] Szeliski, R.: *Computer Vision: Algorithms and Applications*, Springer-Verlag New York, Inc., New York, NY, USA, 1st edition (2010).

[21] Törönen, P., Kolehmainen, M., Wong, G. and Castrén, E.: Analysis of gene expression data using self-organizing maps, {*FEBS*} *Letters*, Vol. 451, No. 2, pp. 142–146 (1999).

[22] Vinh, N. X., Epps, J. and Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary?, *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, New York, NY, USA, ACM, pp. 1073–1080 (2009).